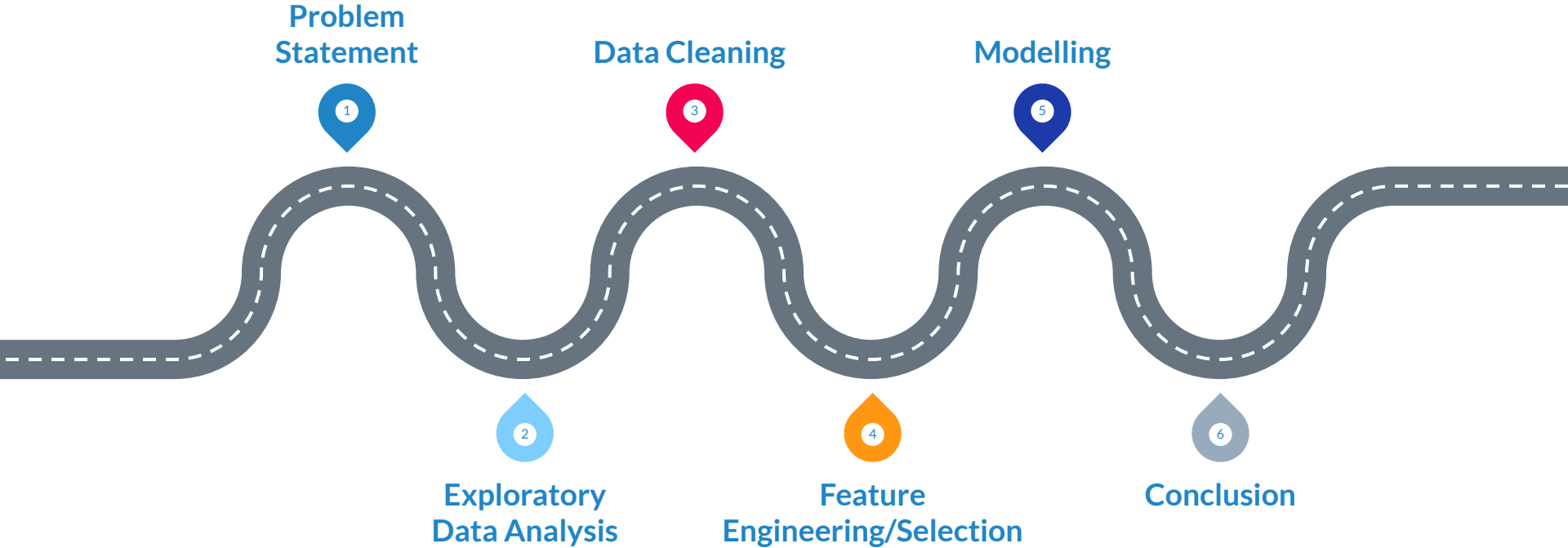# Using Analytics in Real Estate

Group 5 Real Estate Tech Start Up

Members: Benjamin, Anabel, Ahmad, Jeryll

Roadmap

Problem Statement

1

Data Cleaning

3

Modelling

5

2

Exploratory Data Analysis

4

Feature Engineering/Selection

6

Conclusion

Problem Statement:

Mismatch expectation in housing price between sellers and buyers resulting in lengthy negotiation before transaction can be closed

The Solution:

Using analytics to provide accurate market prices to users based on past sales transactions to bridge housing price expectations
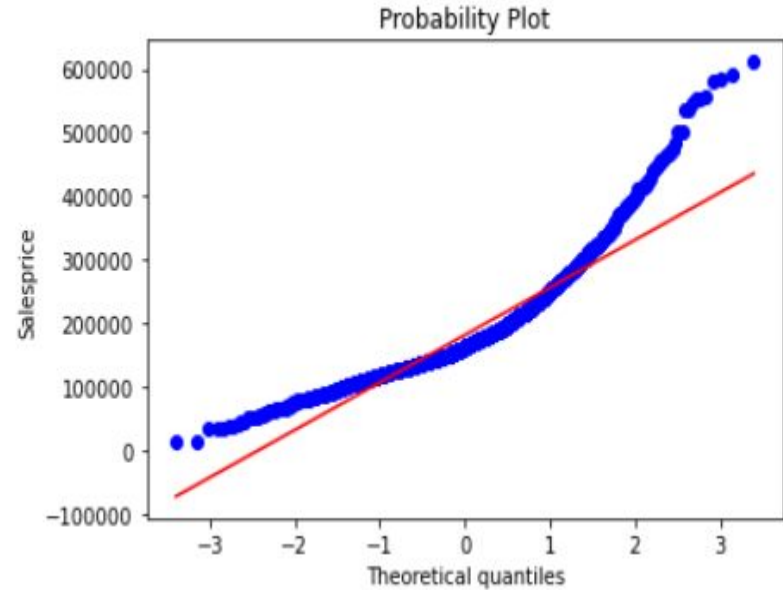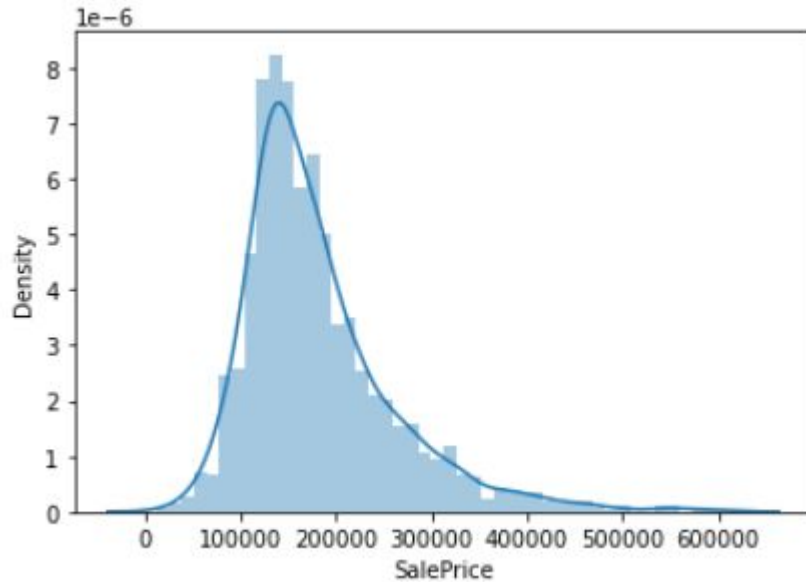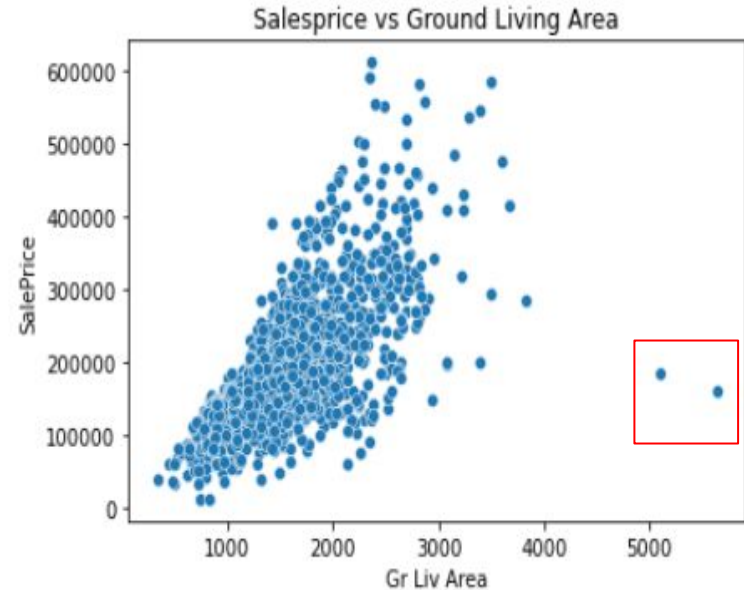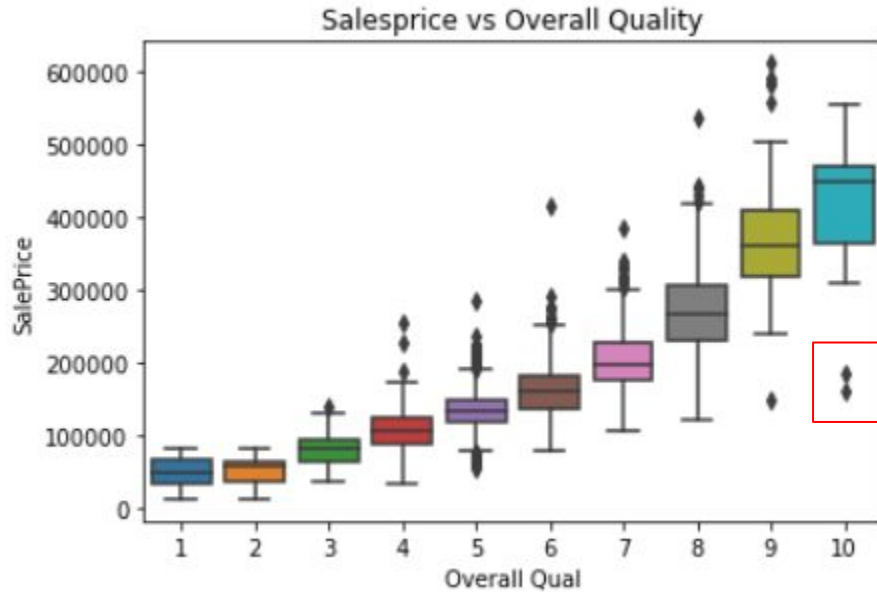
1

Exploratory Data Analysis

- Historical housing price distribution in Ames

- Key Features affecting housing price

Housing Price is positively skewed and do not have linear relationship

There are 2 extreme outliers in the data

Salesprice vs Overall Quality
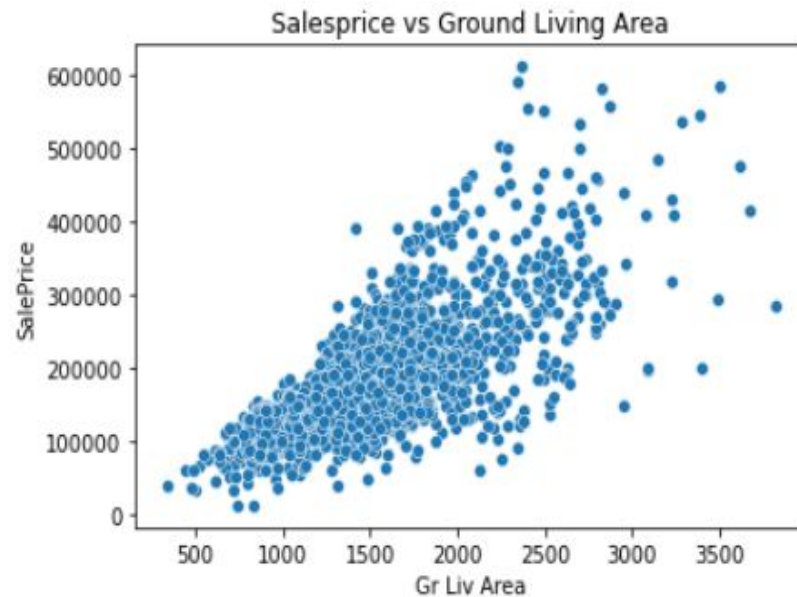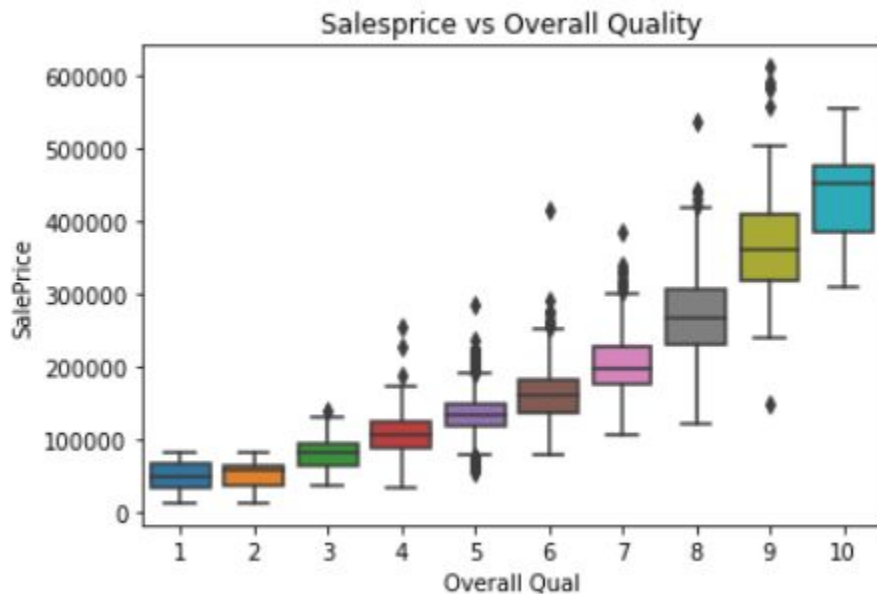
Salesprice vs Ground Living Area

Data Cleaning

- Removing outlier

- Replacing all null values and label encoding ordinal features

- Changing data types to ensure all variables data types are accurate

3

Removing 2 extreme outlier previously identified

Salesprice vs Overall Quality

Salesprice vs Ground Living Area

Data cleaning steps are broken down by the data type of the features

Continuous

- **Replace with 0:** Represent absence of the feature in the house

- **Lot Frontage:** replace null value with the median of the lot frontage by neighbourhood

Ordinal

- **Replace with 0:** Represent absence of the feature in the house

- Label encoding for all ordinal variable e.g. Garage Qual

Nominal

- **Replace with N/A:** Represent absence of the feature in the house

- Change 'MS Subclass' variable to string as this is a nominal variable

Feature Engineering and Selection

- Features elimination

- Adding new features

- Transform sales price to be normally distributed

Feature engineering helps to improve model accuracy

Features elimination

- Removed features with more than 50% null values and does not add significant value to the model

- Removed irrelevant features such as ID, PID and month sold

Adding new features

- Sum up features to create new features such as porch area, total living area and total bathroom (Dropped the original features)

- Create interactions between features which have high correlation with sale price (eg. overall quality * overall area)
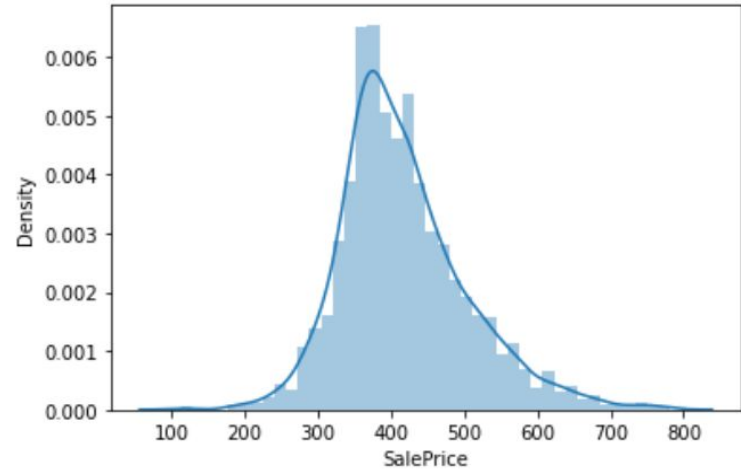
4

Sales price distribution appears more normally distributed after square root transformation

<u>Before</u>

<u>After</u>

4

13
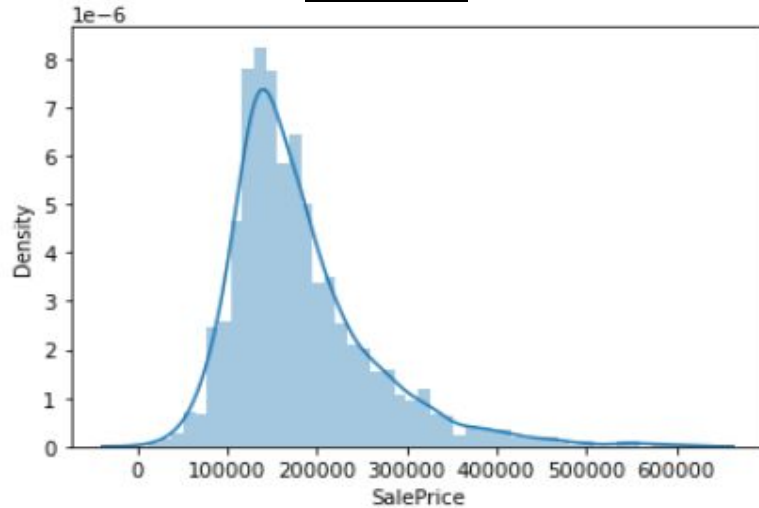
Modelling

- Basic Model with 3 features

- Full Model with all features

Steps before running the model

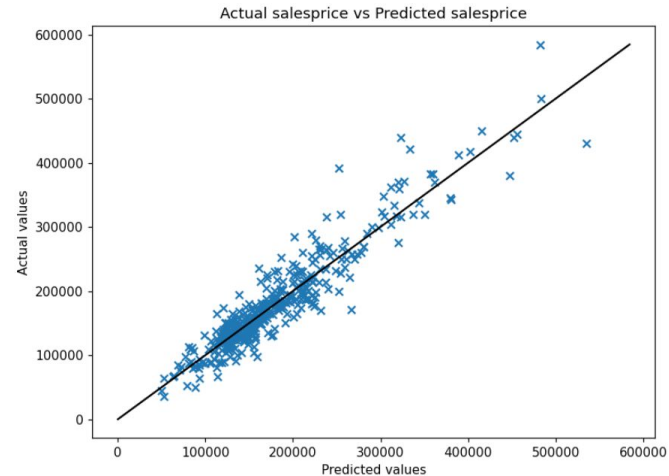| Dummy variables	Train Test Split	Scaling
Create dummy variables for categorical features	Perform train test split with 80% train, 20% test	Scaling the features

5

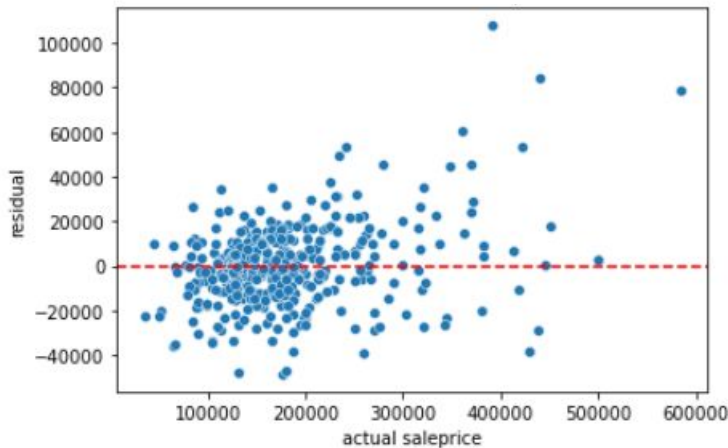# Basic Model: Residual and prediction plot shows residual is randomly distributed around zero

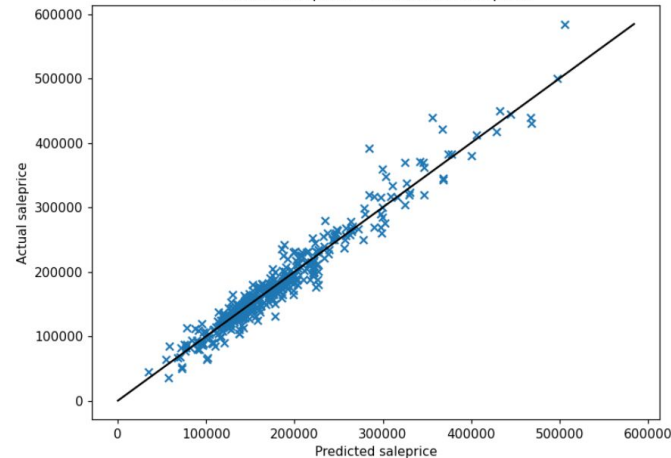**Features:** Overall Condition, Year Built, overall_quality*overall_area

Full Model: Residual and prediction plot shows better results than the basic model

Residual vs Actual Saleprice

Actual salesprice vs Predicted salesprice

Conclusion:

Users can just input 3 compulsory features of the house to obtain the predicted housing price. If users wanted a more accurate prediction, they can input other optional features into the application.

Thank You!

Q & A